DISCUSSION OF PROFESSOR BARNARD'S PAPER

"A COHERENT VIEW OF STATISTICAL INFERENCE"

by

David V. Hinkley
University of Minnesota

Technical Report No. 395
November, 1981

1. Introduction

Professor Barnard has provided a stimulating account of his views on inference. I share his optimism about the possibility of consensus, and I agree with all of his prefatorial recommendations.(*) Nevertheless, I must stress the value of work in areas outside the orbit of Barnard's discussion of the issues if a real consensus is to be achieved. My detailed discussion will focus on four areas which should be addressed in a broadly-based discussion: ancillarity, approximate likelihood methods, nonparametric analysis, and simultaneous estimation. First, however, I should like to mention some points that need fuller discussion in the literature:

(a) the sources and interpretations of probability,

(b) the interaction between modelling and analysis,

(c) extensions of Jeffreys's rational Bayesian theory, and

(d) the extent to which decision theory has provided useful results for inference and methodology.

This last is important because the usual negative attitude of "inferentialists" toward decision theorists, and vice versa, is counterproductive and could preclude a genuine consensus.

2. Ancillarity

The concept and use of ancillarity arise in Barnard's paper, but in a fairly limited setting that gives a rather narrow view of this fundamental idea. I have recently discussed some of the issues in print (Hinkley, 1981), and so I shall try to be brief here.

In statistical inference we deal with probability summaries, and in order that these be as meaningful as possible for the case at hand

(*) see Appendix.

these probabilities should be defined on a relevant set of possible sample outcomes. Such a relevant set would be the smallest set, surrounding the observed outcome, on which a sufficiently rich probability distribution is defined. At the same time, restriction from the prior set of possible outcomes to the actual reference set should not lose information relevant to the inference.

This qualitative notion is made precise in connection with parametric statistical models by defining an ancillary statistic, which indexes the relevant set of outcomes. Thus, if S is minimal sufficient for θ such that $S = (T, A)$, where A's distribution is independent of θ, then A is ancillary. The relevant set of sample outcomes is $\{S: A = a\}$ when a is the observed value of A. There are, however, two kinds of ancillary statistics, namely experimental and mathematical (Kalbfeisch, 1975). The former, A_e, determines a series of performable experiments, and for pedagogical purpose experimental ancillaries are most effective; typical being the "choice of instruments" example (Efron & Hinkley, 1978, Section 1) and the regression example where A_e is the set of explanatory variables. Conditioning inference on A_e is simply equivalent to analysing the particular experiment actually performed. A mathematical ancillary statistic A_m is more difficult, because it is an abstract artefact of the particular stochastic model and seems to have no physically attractive interpretation. However, we must remember that in the interplay between modelling and analysis, A_m provides measures of agreement between data and model, so that conditioning on A_m is equivalent to making an automatic allowance for pre-inference tests of fit. Classical unconditional inference is suspect because it fails to make such an allowance, which seems

surprising in view of the vast literature on goodness of fit procedures such as tests of normality.

Exact ancillarity is a rather restrictive property. Recently the notion of approximate ancillarity has been discussed in connection with large sample likelihood methods; see section 3 of my comments. The general idea is that the information content of A should be o(1), preferably $O(N^{-1})$, as some notional quantity N (such as sample size) increases. This is useful either when an exact ancillary does not exist, or when an exact ancillary defines a singleton reference set. Of course, approximations should be viewed with caution, and should be tested numerically. But at the same time we should remember that in practice all of statistics deals in approximations: exactness is convenient for logical development, but should not be an impediment to application.

The possibility of ancillarity arises also in connexion with randomization and experimental design, where some would argue that the outcome of a randomly selected design is ancillary. Such a fallacy is easy to arrive at if the earlier qualitative explanation is ignored. A legitimate question of ancillarity arises if a design outcome belongs to a particular subset of designs, such as the subsets of diagonal and Knight's Move Latin Squares. For further, albeit brief, discussion, see Hinkley (1980). A thorough treatment of ancillarity should include discussion of this area.

3. Awkward Asymptotics

I would agree that the asymptotics literature is awkward, and surely Barnard is right to cast a critical eye on the work of LeCam, Berkson and others relating to the MLE. However, I believe that much

progress has been made in the last ten years on approximate likelihood methods, and has finally led to constructive understanding of Fisher's work.

There has been a basic misunderstanding between decision theorists and inferentialists. The former solve decision problems (in the abstract, with a monotonous predilection for mean squared error loss), and so they work beyond the inference stage. Fisher gave the basic ingredients of inference, with the MLE as the first step in an approximate sufficiency reduction. Nobody would claim that the MLE itself would be the solution to a specific decision problem. A good discussion of this issue is contained in Efron's 1981 Wald Lecture III.

In the approximate inferential solution of parametric problems, the MLE $\hat{\theta}$ is complemented by successive likelihood derivatives $I = -\ddot{\ell}_{\hat{\theta}}$, $\dddot{\ell}_{\hat{\theta}}$, etc. In effect one uses as many derivatives as would be needed to give an accurate reconstruction of the likelihood itself, so that one keeps a set of statistics which is approximately sufficient. Simplicity will sometimes require parameter transformation, as is well illustrated in work on non-linear regression by Hamilton, Watts, and Bates (1981, 1982). Interest in approximate likelihood methods was kept alive by work of D. Sprott, and in recent years there have been major advances in approximate conditional inference by Amari, Barndorff-Nielsen, Cox, Efron, Hinkley, and the Skovgaards. I would welcome Barnard's comments on these developments.

One of the simplest results that was first emphasised by Efron and myself is that $Q = I^{\frac{1}{2}}(\hat{\theta} - \theta)$ is approximately pivotal and approximately N(0, 1) independent of approximate ancillaries. At the time we worked with i.i.d. variables, but there is evidence now

that Q is approximately pivotal in great generality. As an illustration 1 take an example mentioned to me by D. Siegmund. Let $\{X(s): 0 \leq s \leq n\}$ be the continuous analog of an first-order autogressive process (AR1), defined by

$$dX(s) = \theta X(s)ds + dZ(s), \ \{Z(\cdot)\} \text{ Gaussian white noise.}$$

This process is non-stationary, and the usual unconditional asymptotic result, i.e. $\{E(-\ddot{\ell}_\theta)\}^{\frac{1}{2}}(\hat{\theta} - \theta) \approx N(0, 1)$ as n inceases, fails—as it does in the AR1 process with regression coefficient $\theta \pm 1$ outside (-1, 1). But the pivot Q is exactly N(0, 1). This is proved by showing that

$$\hat{\theta} - \theta = \int_0^n X(s)dZ(s) / \int_0^n X^2(s)ds$$

is exactly $N(0, I^{-1})$ conditional on $I = \int_0^n X^2(s)ds$.

A more useful statement of asymptotic normality would seem to be: $I^{\frac{1}{2}}(\hat{\theta} - \theta) \approx N(0, 1)$, with $I/E(-\ddot{\ell}_\theta) \approx 1$ in ergodic problems. If both results hold, then we have the standard textbook result $\{E(-\ddot{\ell}_\theta)\}^{\frac{1}{2}}(\hat{\theta} - \theta) \approx N(0, 1)$.

Of course when the likelihood is not itself approximately normal in shape, then the approximate inference solution $I^{\frac{1}{2}}(\hat{\theta} - \theta) \approx N(0, 1)$ is not adequate. More refined approximations are discussed by Bandorff-Nielsen, Cox, Hinkley, and Sprott in recent work.

4. Modelling, Sampling and Nonparametric Analysis

Barnard correctly argues that serious efforts should be made to create catalogs of information about models from data. There is no substitute for a correct model. Note, however, that continued monitoring of data will often give information about distributions of parameters. This would suggest that when models are based on sequential experience with similar data sets, (empirical) Bayesian inference will be appropriate.

One area where models do not seem to play much of a role is simple survey analysis. In fact there is a gulf between sampling methods and parametric model analysis in most expositions, perhaps because the familiar exactness of theory in the parametric context is not available for finite population sampling.

Does pivotal inference have a role in sample survey analysis? Suppose that we are interested in a population mean, and sample $n = 20$ items from a population of 1000. If we know nothing about the population, should we base our inference on $(\bar{X} - \mu)/s$, or on $\bar{X}/\mu$? Whichever we choose, how should we use it knowing that normal approximations are often inadequate? There are few answers to be found about this very basic problem in the literature. One useful approach suggested by Efron (1981) is to estimate the distribution of a chosen "pivot," such as $(\bar{X} - \mu)/s$, by simulation with samples drawn with replacement from a population whose frequency distribution is uniform on the observed data values. This simple idea is often effective.

The situation is even more difficult if we wish to compare two finite population means. The following unusual suggestion is based on an idea of Efron's. Suppose we have two independent samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ and that we wish to compare the two population means m_x and m_y. Further, take the extreme position of not assuming that values other than those observed are possible. Then our populations are modelled as follows

	Population 1	Population 2
possible values	$x_1, \ldots, x_n$	$y_1, \ldots, y_n$
frequences	$f_1, \ldots, f_n$	$g_1, \ldots, g_n$

The MLE's of f_i and g_i are $\hat{f}_i \equiv n^{-1}$, $\hat{g}_j \equiv n^{-1}$. The null hypothesis of equal means, $m_x = m_y$, forces the constraint $\Sigma x_i f_i = \Sigma y_j g_j$. Suppose that we wish to assess the signicance of an observed positive difference $\bar{x} - \bar{y}$. Following the traditional conservative approach, we calculate significance probability under the specific null model which makes the data as plausible as possible—interpreted here to mean the specific frequency vectors $\underset{\sim}{f}$ and $\underset{\sim}{g}$ which minimize $\Sigma f_j \log(\hat{f}_j/f_j) + \Sigma g_i \log(\hat{g}_i/g_i)$. This gives the simple null model $f_j \propto \exp(\lambda_0 | x_j), g_i \propto \exp(-\lambda_0 | y_i)$ with λ_0 satisfying $E_{\lambda_0}(X - Y) = 0$. In the family parametrized by λ, the unconstrained MLE corresponds to $\hat{\lambda} = 0$. The required significance probability

$$P = \text{prob}_{\lambda 0}(\bar{X} - \bar{Y} \geq \bar{x} - \bar{y})$$

can be expresses as

$$P = \sum_{s=x-y}^{\infty} \underline{e}\exp(n\lambda_0 s)\ \text{prob}_{\hat{\lambda}}(\bar{X} - \bar{Y} = s),$$

where of course $\hat{\lambda} = 0$. Thus the significance probability can be estimated by simulation under $\lambda = 0$, i.e. using repeated samples drawn with replacement from the original data.

This proposed procedure is extreme, in the sense that absolutely no modelling is involved. In the general context of Boostrap methods, it is but one of many possible procedures. I would be interested to know Barnard's thoughts on this kind of problem, and in particular his thoughts on the relevance of pivotal inference.

5. Simultaneous Estimation

There is a brief mention of this topic in Barnard's paper, where he suggests that shrinkage estimates are unethical. I think this comment was ill-advised and based on some misconceptions about

simultaneous estimation methods. Perhaps this is because much of the post-Stein work was pretty sterile decision theory, and more recently Stein's ideas have been perverted in the naive work on ridge regression. I would bet very heavily that if any consensus is reached in statistics, then the consequences of Stein's work will have a prominent place. We should try to understand in a qualitative sense what Stein's mathematical results are about—this kind of understanding led to important practical work by Efron and Morris, in an empirical Bayes mould. Roughly speaking, if the data strongly suggest a distribution for parameters, then we have a situation where Bayesian pivotal inference applies: in the simple case of estimating means $\mu_i = \bar{x}_i + s_i e_i$ and $\mu_i = m + t_i f_i$ where s_i and t_i are scale statistics, m is an overall mean, and e_i, f_i are independent with known pivotal distributions, why would it be unethical to use all of the information? Not to do so in Barnard's census application might lead to rewards for sloppy surveys which give some unbiased estimates the advantage of extreme sampling errors.

Of course one should pay attention to questions of robustness, and attempt proper modelling of the application (unlike what happens in ridge regression). More recent work by Stein has done this; see Stein (1981).

Perhaps the language of decision theory is at fault in our under-appreciation of the work on simultaneous estimation. This is why I think that we should seriously assess the practical implications of decision theory—it is not useless, but may seem to be simply because it is in a language foreign to many of us.

REFERENCES

Bates, D. M. and Watts, D. G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann. Statist., 9, 1152-1167.

Efron, B. (1981). The jackknife, the bootstrap, and the resampling plans. Stanford University Biostatistics Report No. 63.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika, 65, 457-482.

Hamilton, D. C., Watts, D. G., and Bates, D. M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. Ann. Statist., 10 (to appear).

Hinkley, D. V. (1981). Likelihood. Can. J. Statist., 8, 151-163.

Hinkley, D. V. (1980). Comments on paper by D. Basu. J. Amer. Statist. Ass., 75, 582-584.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality (with discussion). Biometrika, 62, 251-268.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. Ann. Statist., 9, 1135-1151.

APPENDIX

Preface to "A Coherent View of Statistical Inference"

The following pages present arguments in favour of a number of changes in statistical practice. Other arguments can be made leading to similar recommendations. It is more important for us statisticians to agree on our practical recommendations than on the reasons underlying these. I therefore have listed the principal recommendations here, particularly to invite your judgement, of support or of opposition:

(i) We should discourage statements of the form 'P < .05' in relation to testing hypotheses. There can be little excuse, if any, nowadays, for omitting to specify P = so and so. It can then be for the reader to choose his level of significance, if he wants one.

(ii) The P value should be conditional on the occurrence of any aspect of the data whose probability would not be changed by abandoning the hypothesis H_0 tested.

(iii) In addition to quoting the P value, whenever it is possible to formulate a reasonably plausible alternative to H_0, the likelihood ratio should be quoted in addition to the P value. If there is a family $H(\theta)$ of alternatives, a suitably weighted compound of $H(\theta)$ may be used for this purpose, or alternatively the likelihood function can be plotted. The P value and the likelihood ratio are each measures of strength of evidence, neither being superior to the other, in general.

(iv) We should abandon the notion of estimation as attempting to find a single number 'closest', in some sense, to the true parameter value. The object of estimation is to make an estimation statement, one of the simplest forms of which would be

$$\theta = t(x) + s(-e) \qquad (e \text{ density } \psi(e))$$

which is to be interpreted as meaning that $(t-\theta)/s$ has the distribution specified by the density ψ. This last should be in some standard, agreed form, for instance having mode or median at 0 and a unit semi-interquartile range. More generally, the estimation statement may have to take the form

$$f(x,\theta) \text{ has density } \psi$$

where, for given x, the mapping $f(x,\theta)$ is 1-1. Whether such statements are possible may depend on the particular data to hand.

(v) Because the conditional argument allows us to treat, without undue difficulty, arbitrary observations densities, we should cease trying to fit data analyses to the Procrustean bed of normality of distribution, and instead gather empirical data on distributional forms and use these in our analyses.

(vi) We should recognize that 'robustness' of inference is a conditional property--some inferences from some samples are robust, in the sense that changing assumptions about distributional form matters little. But other inferences, or the same inferences from other samples, may depend strongly on the distributional assumptions. In the latter case it is the statistician's job to point to the dependence on distributional form, <u>not</u> to obscure this fact. This is part of the general duty to tell clients not only what they may take as known, but also what they should take as unknown, in so far as it may be relevant to the matters they are concerned with.

(vii) We should recognize that probability is a complex concept, in its application to data, and that it is too much to hope that all questions can be answered in terms of P values, likelihood ratios, or estimation statements of a simple kind. More complex forms of final inference,

for example such as that illustrated in the text in connection with the Behrens-Fisher probolem, may be needed. We have put ourselves in shackles by attempting to keep things over-simple; it is time we broke out.

(viii) The Bayesian controversy is irrelevant to statistical inference. Since the method of analysis is the same, whether or not prior distributions are or are not admitted, the question becomes whether or not a given assumption is justifiable in each given case.